Discussion on *Bob Calfee's February, 2009 ITQ Panel Presentation and Notes*

Submitted by PJ Hallam, ITQ A&D Member

March 25, 2009

On February 6, 2009, Improving Teacher Quality (ITQ) Assessment and Dissemination (A&D) Team members PJ Hallam and Don Hubbard convened a panel discussion on *Day-to-Day Lessons of Researching Teacher Professional Development* at their annual project meeting. One of the speakers, Dr. Bob Calfee, generously shared his notes with the ITQ community after the panel discussion.

Sage has just published a compilation of essays that covers these points, and others, in more detail, *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* edited by Donaldson, Christie, and Mark. ITQ A&D team members recommend this book.

Before the book was published, PJ and two Research Directors, Doug Grove and John Gargani, had an email conversation about a few of the points in Dr. Calfee's (Bob) notes.

Howard Levine, also a member of the A&D team, Don, and I thought this conversation was worthy of sharing with other ITQ members. From my perspective, Doug and John "unpacked" meaning from Bob's brief notes, enhancing my understanding, and extending the concepts.
If you wish, after reading, please feel free to send me your comments at pjhallam@speakeasy.org.
Enjoy.

P. J. Hallam (PJH): *Bob brings up the importance of statistical power and how it's often overlooked. Since I've nagged my project's RDs about including power in their reporting and analyses, I felt validated!*

Doug Grove (DG): Power should be a part of the reporting and most researchers doing experimental or quasi-experimental studies most likely estimated power effects when they wrote the evaluation design to the RFP. Most RFP's require the inclusion of an effect size and this might be something CPEC could make standard in the RFP. I don't remember if the RFP required effect size or not (we included it).

John Gargani (JG): Power is very important as a planning tool, never useful as a post hoc explanation. A power analysis conducted before a study gives researchers some assurance that when the data are collected and submitted to the specified statistical test it will provide a useful answer to their question. If power looks OK, it means that one possible mode of the evaluation's failure has likely been avoided. If it does not, it means that the study's failure is very likely. Having collected the data, performed the statistical test, and concluded that you cannot reject the null hypothesis, there is no value in conducting a power analysis—the null hypothesis is always false at some level of precision, so if you had more power you would have found a statistically significant difference. This is why, as Doug mentioned, results should be reported in a number of ways—estimates in the original units (usually the most easily interpretable), the results of the statistical test (p-values, standard errors, t-values, and the like), estimates as standardized effect sizes, and the relationship of the standardized effect sizes to the researchers' notion of what constitutes a practically meaningful amount and/or the results of prior research. Having said that, researchers should identify in advance what they deem to be practically significant and justify it, and they should use Cohen's rule of thumb regarding effect sizes with caution because it is usually misinterpreted (big and small relate to the relative ease of detection, not historically obtained results).

PJH: *Bob mentions treating teachers as individuals. At first I thought he meant that he preferred this approach, but I think he's making the case that the nested nature of students in classes in* *schools is important –e.g., Clotfelter (2006), school effects on teachers—and in the processes is supporting the use of HLM. I'm a big fan of HLM, but it's not always possible.*

DG: HLM is an excellent analysis method, *if* the intervention and hypothesized effects are best analyzed via HLM. However, HLM is currently treated like a magic wand that a lot of people are using. Bob may be cautioning against HLM becoming a default way to analyze data in the hopes of finding something, anything nested somewhere. Again, I think HLM is great if assumptions are met and the data collected lends itself to HLM analysis.

JG: I like to keep three units in mind. The randomized units are the people, groups, organizations, families, etc. (things) that we place into treatment and control/comparison groups. Measured units are the things we measure to gauge outcomes. The treated units are the things that experience (or could have experienced) the treatment. So we could randomize schools, treat teachers, and measure students. Or we could randomize students, treat teachers, and measure schools. Or…you get the picture. You choose the units based on the research questions you have, your theoretical rationale for why the program will work, and pragmatic reasons such as the immediate context, budget, and time. The relationship of these three units, especially if you measure repeatedly over time, can take many forms and be extremely complicated. In any event, whenever these three units are not the same, we introduce complications of some sort that we need to plan for in advance if we expect the evaluation to provide useful information. You can address some of the units-mismatch issues that I described (like nesting) with HLM and other related models. In many cases, HLM is a necessity because it offers the most appropriate statistical tests related to program impact estimates.

PJH: *Are very large numbers needed for HLM? Can as few as 25 teachers make for a strong study?*

DG: The relatively small number of teachers in some studies makes it difficult to look for effects using HLM, and the size of the sample minimizes generalizability for sure. Small sample sizes can work in quasi-experiments, but these designs can be drastically compromised by attrition in either the experimental or comparison groups. It depends on the unit of analysis. I am currently evaluating a Teaching American History grant and a key goal of the grant is to look for an increase in teacher knowledge of American history. The teacher has to be the individual examined to meet that goal and we have set up a quasi-experiment to determine if participating teachers increase their knowledge of American history more than those not in the program. HLM does require some larger sample sizes that more traditional methods of analysis. Student numbers are usually not the problem. It is the school-level and teacher-level analysis that suffers from inadequate sample sizes. 25 teachers would probably be more appropriate to a mixed methods study or a very tightly designed quasi-experiment. Mixed methods studies can produce very important findings on implementation and effects of professional development. Problem with these mixed methods studies is they are not recognized by Institute for Educational Science or "What works Clearinghouse" as rigorous scientific research. Personally, I have no problem with mixed methods.

JG: The rule of thumb is "you analyze like you randomize," meaning that if you randomize teachers then you compute a treatment effect at the teacher level. In this case, if you are using student test scores as outcome measures, then principally it is the number of teachers that drives statistical power. The number of teachers you need depends on several factors, but the simple answer is that 25 teachers can sometimes be enough. I wrote a short brief on this with Tom Cook, though in that case we were looking at schools rather than teachers, but the lesson is the same -- when conditions are right you can reasonably conduct "small" studies that violate the prevailing wisdom that HLM always requires "large" samples of say 40, 50, or 60 teachers. Nevertheless, if a researcher cannot demonstrate this with a power analysis before the study begins, why believe that this is one of those instances? There is a lot to this, but as Doug mentioned one thing is certain – you will always need more students for a teacher-level or school-level HLM study than you would if you did a student-level study.

PJH: *Isn't it possible that most readers could understand HLM results without necessarily understanding the entire process? It provides impact estimates, p-values and correlation coefficients just like other regression analyses. Readers who are qual snobs would have excluded themselves already; wouldn't be losing additional potential consumers/stakeholders.*

DG: Perhaps the larger question is, "Do we have a choice?" NCLB is clear on the kinds of research it considers the gold standard. There have been many articles by Scriven and Colmer on this notion of the gold standard. Many of these articles point to the cost/benefit of these kinds of studies and the difficulty of conducting them in complex educational setting with students, teachers, parents, budget cuts, and all the other variables that need controlled. Mixed methods studies can produce very important findings on implementation and effects of professional development. The problem with these mixed methods studies is they are not recognized by Institute for Educational Science or "What Works Clearinghouse" as rigorous scientific research. Right now the government has been pretty clear on the kind of research it wants done: RCT's, quasi-experiments with tight controls, and regression discontinuity studies.

JG: I agree that if you can understand the results of a t-test, you can understand the impact estimates produced by HLM. Even if that were not true, researchers should use HLM when it is appropriate. Sometimes, as Doug mentioned, it—or perhaps more generally randomized trials and quasi-experiments—can be required when they are arguably inappropriate. This sort of "scientifically-based research" is more difficult and costly than other forms of research, so we should think hard about when and how to undertake it. I believe that it is generally a good idea when three conditions are met. First, the program is sufficiently well developed that it will not be changed substantially in the near future and there is a reasonable expectation that the program can be implemented well and to good effect. Second, the researchers are conducting what might be called a traditional confirmatory analysis, which means that before they start the study they have clearly articulated questions that they have grounds for believing they can answer by submitting precisely specified data to specific statistical tests. Third, the primary question at hand is whether the program offered a particular benefit or benefits to the program participants. This primary question is about the past performance of the whole program and it differs from other similar sounding questions like whether the program works (a question about the future), whether a component of the program like lesson study provided a benefit (a question about a part of the program), or whether it was a good idea to implement the program (a comparison of benefits across all programmatic options). To my mind, these three conditions provide the fundamental rationale for conducting SBR.

PJH: *Both of you describe scientifically based research as challenging and problematic in practice. Final thoughts on what makes scientifically-based research so difficult?*

DG: One important point that is often overlooked is resources, in that trying to do SBR is really a budget killer. These kinds of studies, or even mixed methods studies, can't be properly done on 20% of the total grant award. My suggestion is that future reviews of the new RFPs make sure that the type of research being proposed is adequately funded.

JG: Doug makes a good point. While I am a frequent critic of what I consider to be inefficiency in evaluation and research, the time and attention it takes to observe classrooms, recruit teachers, develop new tests and surveys, and publish results is tremendous. With inadequate budgets, the failure rate of studies goes up while the information content of studies goes down. Given the current level of funding, the wise move would be either to trim expectations or increase research budgets.